BIOINFORMATICS

Vol. 19 no. 13 2003, pages 1664–1671
DOI: 10.1093/bioinformatics/btg204

PRECIS: Protein reports engineered from concise information in SWISS-PROT

A.L. Mitchell[1,2,*], J.R. Reich[3] and T.K. Attwood[1,2]

[1]School of Biological Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, UK, [2]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [3]Discovery Informatics, Basilea Pharmaceutica Ltd, PO Box 3255, CH-4002, Basel, Switzerland

Received on October 31, 2002; revised on February 25, 2003; accepted on March 5, 2003

ABSTRACT

Motivation: There have been several endeavours to address the problem of annotating sequence data computationally, but the task is non-trivial and few tools have emerged that gather useful information on a given sequence, or set of sequences, in a simple and convenient manner. As more genome projects bear fruit, the mass of uncharacterized sequence data accumulating in public repositories grows ever larger. There is thus a pressing need for tools to support the process of automatic analysis and annotation of newly determined sequences. With this in mind, we have developed PRECIS, which automatically creates protein reports from sets of SWISS-PROT entries, collating results into structured reports, detailing known biological and medical information, literature and database cross-references, and relevant keywords.

Availability: The software is accessible online at: http://www.bioinf.man.ac.uk/cgi-bin/dbbrowser/precis/blast_precis.cgi

Contact: mitchell@ebi.ac.uk

INTRODUCTION

The fecundity of genome projects and the growing number of uncharacterized sequences in public databases has created a demand for methods to annotate sequences automatically. Computational methods to unearth relationships between sequences tend to rely on two main techniques. One involves standard pairwise similarity searches, e.g. using FASTA (Pearson and Lipman, 1988) or BLAST (Altschul *et al*., 1990). The principle here is that the top-scoring match is likely to be a close relative, and therefore any annotation associated with it can be inherited by the query. The other involves seeking similarities between the query and diagnostic 'patterns' housed in protein family databases, such as PROSITE (Falquet *et al*., 2002), PRINTS (Attwood *et al*., 2002), Pfam (Bateman *et al*., 2002) and InterPro (Apweiler *et al*., 2001). The idea here is that the best match will be indicative of the evolutionary family to which the query sequence belongs, and functional and structural annotation can be inherited accordingly. Used wisely, these approaches can be used to gain biological insights into uncharacterized sequences. Sometimes, however, the top hit may not be the best biological match, and erroneous annotation may consequently be attached to the query.

To address such problems, 'expert systems' have been developed that combine pairwise and family database searches: e.g. GeneQuiz (Scharf *et al*., 1994), MAGPIE (Gaasterland and Sensen, 1996), and PEDANT (Frishman and Mewes, 1997). Nevertheless, even in these systems, function assignment tends to rest with the best FASTA or BLAST hit. Another approach attempts to facilitate the inference of protein function by exploiting the semantic differences inherent in sequence and family database search outputs (Selley *et al*., 2001), digesting the results of multiple searches and providing a consensus diagnosis of the best match, placed in the context of the family to which it belongs. Other approaches have turned to the literature, extracting keywords from MEDLINE abstracts gathered for families of related sequences and comparing them with those of unrelated families. Methods of this type have been incorporated into tools such as GeneQuiz (e.g. Andrade and Valencia, 1998), but are not generally available. In another technique, keywords are distilled directly from SWISS-PROT, taking advantage of its structured annotation, and of the controlled vocabularies and syntax used to populate its different fields (Wise, 2000). Although SWISS-PROT is more amenable to this type of analysis than the free texts in MEDLINE, the results are nevertheless somewhat disappointing. Ultimately, keywords are of little practical value for an annotator whose task it is to write about the biology and medical significance of families of sequences.

To address many of these issues, we have developed a new annotation tool, PRECIS. PRECIS focuses on SWISS-PROT (pilot studies showed literature abstracts to be relatively information-poor with respect to protein families) and, moving beyond mere keyword lists, it attempts to offer

*To whom correspondence should be addressed.

```
ID   OPSD_ANGAN     STANDARD;      PRT;   352 AA.
AC   Q90214;
DT   01-NOV-1997 (Rel. 35, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Rhodopsin, deep-sea form.
OS   Anguilla anguilla (European freshwater eel).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Actinopterygii; Neopterygii; Teleostei; Anguilliformes; Anguilloidei;
OC   Anguillidae; Anguilla.
OX   NCBI_TaxID=7936;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Retina;
RX   MEDLINE=96156843; PubMed=8587887;
RA   Archer S.N., Hope A., Partridge J.C.;
RT   "The molecular basis for the green-blue sensitivity shift in the rod
RT   visual pigments of the European eel.";
RL   Proc. R. Soc. Lond., B, Biol. Sci. 262:289-295(1995).
CC   -!- FUNCTION: VISUAL PIGMENTS ARE THE LIGHT-ABSORBING MOLECULES THAT
CC       MEDIATE VISION. THEY CONSIST OF AN APOPROTEIN, OPSIN, COVALENTLY
CC       LINKED TO CIS-RETINAL.
CC   -!- SUBCELLULAR LOCATION: INTEGRAL MEMBRANE PROTEIN.
CC   -!- TISSUE SPECIFICITY: ROD SHAPED PHOTORECEPTOR CELLS WHICH MEDIATES
CC       VISION IN DIM LIGHT.
CC   -!- DEVELOPMENTAL STAGE: WHEN EEL MATURES SEXUALLY AND MIGRATES BACK
CC       TO DEEP SEA BREEDING GROUNDS THE VISUAL PIGMENTS IN ITS ROD
CC       PHOTORECEPTORS CHANGE FROM BEING MAXIMALLY SENSITIVE TO GREEN
CC       LIGHT TO BEING MAXIMALLY SENSITIVE TO BLUE LIGHT. IN PART, THIS
CC       CHANGE IN SENSITIVITY IS DUE TO A CHANGE IN THE OPSIN COMPONENT OF
CC       THE VISUAL PIGMENT MOLECULE; THIS BLUE SENSITIVE RHODOPSIN IS
CC       EXPRESSED DURING LIFE IN BLUER OCEANIC WATERS.
CC   -!- PTM: SOME OR ALL OF THE CARBOXYL-TERMINAL SER OR THR RESIDUES MAY
CC       BE PHOSPHORYLATED.
CC   -!- MISCELLANEOUS: THIS OPSIN HAS AN ABSORPTION MAXIMUM AROUND 482 NM.
CC   -!- SIMILARITY: BELONGS TO FAMILY 1 OF G-PROTEIN COUPLED RECEPTORS.
CC       OPSIN SUBFAMILY.
DR   EMBL; L78008; AAA99297.1; -.
DR   HSSP; P02699; 1BOJ.
DR   GCRDb; GCR_1248; -.
DR   InterPro; IPR000276; GPCR_Rhodpsn.
DR   InterPro; IPR001760; Opsin.
DR   Pfam; PF00001; 7tm_1; 1.
DR   PRINTS; PR00237; GPCRRHODOPSN.
DR   PROSITE; PS00237; G_PROTEIN_RECEP_F1_1; 1.
DR   PROSITE; PS50262; G_PROTEIN_RECEP_F1_2; 1.
DR   PROSITE; PS00238; OPSIN; 1.
KW   Photoreceptor; Retinal protein; Transmembrane; Glycoprotein; Vision;
KW   Phosphorylation; Lipoprotein; Palmitate; G-protein coupled receptor.
```

Fig. 1. Excerpt from a typical SWISS-PROT entry. The 2-character tags (left) indicate the type of information included on a given line: ID gives the database identifier, AC the accession number, DE the protein description, etc. The CC (comment) field is semi-structured, special sub-fields denoting the protein function, subunit arrangement, sub-cellular location, disease associations, family relationships, etc. Bold lines are those used to engineer PRECIS' reports.

comprehensive reports on protein structure, function and disease in a format that is English-like.

SYSTEM AND METHODS

Data collection

PRECIS takes as input a list of SWISS-PROT identifiers (IDs) and retrieves the full database entry for each. SWISS-PROT entries are characterized by two-character tags that indicate the type of data contained on each line, as shown in Figure 1. For a given set of SWISS-PROT entries, PRECIS only collects information from the following tagged lines: ID (identifier); AC (accession number); DE (description); RN, RP, RA, RT and RL (literature reference number, comments, authors, title and location/citation, respectively); CC (comments, structured by topic); DR (database cross-references) and KW (keywords). This ensures that we capture the most relevant textual information and helps to reduce the amount of data to be processed.

To create meaningful reports from the culled information, we must determine whether the collected sequences constitute a gene family or super-family (united by a common

function), or a domain family (containing a common structural motif). For proteins belonging to the same family, we assume their SWISS-PROT entries will contain elements of common annotation relating to function, structure, etc. However, entries of sequences belonging to different families in a super-family, or that include a domain found in different multi-domain proteins, are likely to share only small amounts of annotation, which is unlikely to be function specific. To distinguish these cases, we refer to the collected ID codes.

SWISS-PROT IDs comprise two elements, the first denoting the protein type and the second the species: e.g. PM22_HUMAN is peripheral myelin protein 22 from human. Sequences belonging to families tend to be characterized by homogeneous IDs (e.g. **LMIP**_HUMAN, **LMIP**_BOVIN, etc. belong to the lens fibre membrane intrinsic protein family). IDs in super-families and domains, however, have variable first elements (e.g. MUP_RAT, LACB_BOVIN, RETB_HUMAN, etc. belong to the lipocalin super-family, while UROT_HUMAN, PLMN_PIG, APOA_MACMU, etc. contain kringle domains). The collected IDs are therefore checked for a common root in the first element (e.g. **LMIP**) down to their first two characters. If a majority (>75%) is found, the sequences are assumed to belong to a family. Super-families and domains require more subtle analysis; nevertheless, information is gathered in a similar way.

Annotation formatting

To the information collected from the ID, AC, DE, R(x), CC, DR and KW fields, rules are applied to determine what information is common to all entries, and what is unique but nevertheless important to retain, as summarized in Figures 2 and 3. The report created details: the name or description of the protein; database cross-references where further information might be found; a set of literature references; a description of its function, structure and associated diseases, the family to which it belongs and/or the domains it contains and, finally, a set of keywords.

Title Within a protein family, most members will share the same or similar annotation on the DE line. All DE descriptive terms are therefore ranked according to their frequency of occurrence, and the most frequent is inherited as the report title. If there are no common terms (rare in a collection of sequences belonging to a genuine family, but usual for super-families and domain families), we refer to the CC *Similarity* sub-field (see later).

Database cross-references To obtain links that might provide structural, functional, family or disease-related information, we select a common core of resources from the DR lines [specifically, PRINTS, PROSITE, Pfam, InterPro, PDB (Berman *et al*., 2000) and MIM (Hamosh *et al*., 2002)]. All DR lines are collected, and duplicated links are discarded. If a PDB link is found, additional links are synthesized, namely to SCOP (Lo Conte *et al*., 2002) and CATH (Pearl *et al*., 2001), which are not cross-referenced in SWISS-PROT. If a PDB cross-reference is not available, information is sought from the HSSP (Holm and Sander, 1999) link, on the basis that this should provide a suitable model.

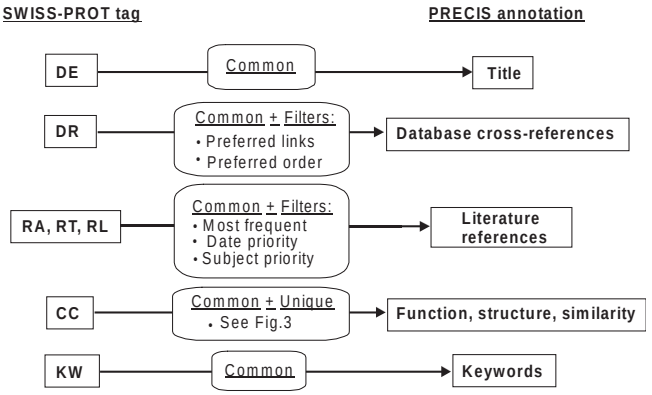

Fig. 2. Illustration of the mapping procedure from specific SWISS-PROT tags to elements of the final report. Common annotation in each field is subjected to different rules and filters to determine which elements are retained. The CC field is the most structured, and provides the core of the bio-medical annotation (see Fig. 3).

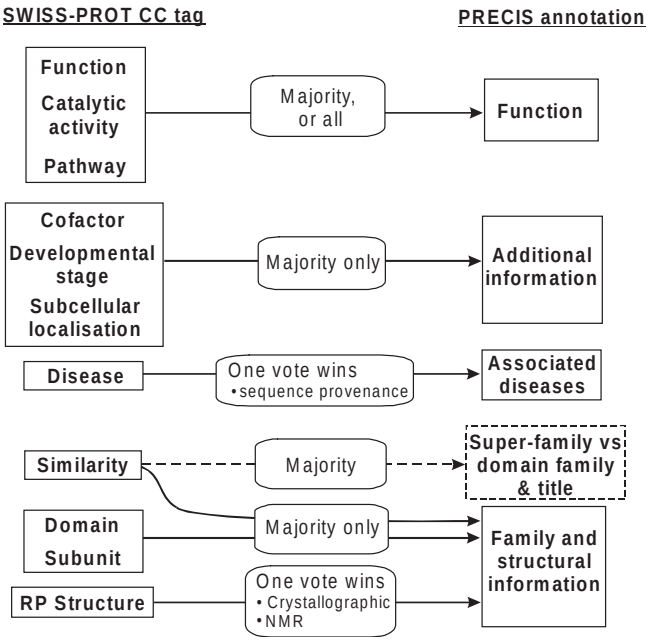

Fig. 3. Illustration of the mapping procedure from SWISS-PROT CC tags to elements of the final report. Annotation in each sub-field is subjected to rules and filters to determine which elements are retained. Structural descriptions are augmented with information derived from RT lines via pointers from corresponding RP lines. Information from the Similarity sub-field is also used to distinguish super-families from domains (and to generate their report titles), as indicated by the dotted line.

Literature references These are gathered automatically from the reference lines [R(x)]. The analysis selects the most frequent shared articles, but if common references are not found, then the most recent, non-shared publications are included. In addition, papers containing details of structure determinations over-ride the requirement for the reference to be shared. RP lines are used to ascertain whether a structure is available. If these lines contain strings such as, 'X-ray crystallography' or 'structure by NMR', the corresponding reference is included in the report.

Function, structure and disease-related annotation SWISS-PROT CC fields are semi-structured and annotation rich. They therefore provide the core of the bio-medical annotation to the final report. The CC field is divided into several sub-fields. Those pertinent to the report include: *Catalytic activity*, *Cofactor*, *Developmental stage*, *Disease*, *Domain*, *Function*, *Pathway*, *Similarity*, *Subcellular location* and *Subunit*. For simplicity, we explicitly ignore sub-fields that are likely to be sequence- rather than family-specific (and therefore are unlikely to convey common information), or that do not relate directly to structure, function and/or disease. However, the system can be easily modified to restore and process any currently ignored sub-field.

Annotation in each selected sub-field is filtered to determine which elements are retained (see Fig. 3). Descriptions in the *Function* sub-fields are inherited if shared by a majority of sequences. If no majority is found, the family is probably functionally diverse, so all function comments are included to reflect this. As with structural references, information relating to disease over-rides the requirement to be common; thus all unique descriptions from the *Disease* sub-field are included, together with an indication of the sequence to which they relate. Structural information is derived from an amalgamation of unique descriptions contained in the *Domain* and *Subunit* sub-fields and information derived from the RT lines via pointers from the corresponding RP lines. A sentence is synthesized using a standard template that states, 'The structure has been determined, e.g. 'Title 1' [*i*] and 'Title 2' [*j*]', where Titles 1 and 2 are the titles of crystallographic and NMR structure determination papers extracted from the RT lines. The bracketed numbers indicate that the papers have been added as *i*th and *j*th articles to the shared papers generated from the earlier reference-gathering process.

Family information and keywords The Similarity sub-field is processed to provide an indication of the family to which the protein belongs or the domains that it contains—here, the most frequently occurring description is used. Finally, all keywords are collected from the KW lines, duplicates are removed, and a non-redundant list is provided.

Domains and super-families versus families

The above discussion relates primarily to the creation of reports for sequences in families. This works relatively well because, generally, information in SWISS-PROT is sequence specific. It is thus simple to gather biological details, with the assumption that they will pertain to the family as a whole. Less straightforward are super-families or domain families, because SWISS-PROT includes little specific information for these cases. Here, modified reports are created, where information is collected in a family-specific manner, and the most highly populated families are taken as representatives of the particular domain- or super-family. The method for dealing with super-families and domains is similar, but there are important differences.

Although protein families may be functionally disparate, they may belong to super-families that share 'high-level' functional and structural similarities. For example, muscarinic acetylcholine receptors modulate physiological functions such as intestinal smooth muscle contractions and heart rate; while opioid receptors mediate analgesia, nausea, euphoria, physical dependence, and so on. Nevertheless, muscarinic and opioid receptors belong to a super-family of proteins that share a common structural framework of seven transmembrane (TM) helices, all of which transduce extracellular signals by coupling to G proteins when activated by their endogenous ligands. By contrast, sequences that share structural domains are unlikely to share high-level functions, and structural similarities will be confined to the specific domain. For example, SH2 domains are small protein modules found in different protein contexts: e.g. in association with catalytic domains of non-receptor protein tyrosine kinases; in structural proteins like tensin; and in a group of small adaptor molecules, such as oncoprotein Crk. It is therefore important to be able to distinguish super-families from domains, otherwise structural and functional information in the final report is likely to be misleading.

As we saw earlier, super-families and domain families generally differ from families in not containing a majority of sequences with the same root ID; they are hard to distinguish, as they have rather flat distributions of disparate IDs. To make the distinction, we refer to the CC *Similarity* sub-field. For protein families and super-families, the word 'family' tends to be used [e.g. 'belongs to family 1 of G protein-coupled receptors' (GPCRs)]. For domains, however, the most commonly occurring word is more likely to be that of the domain (e.g. 'contains 2 SH2 domains'). For super-families and domains, the most frequently occurring informative terms in this sub-field are therefore used to provide the report title, rather than terms in the DE lines.

For super-families, PRECIS produces a report detailing up to five of the most highly populated families and, for each, digests the CC sub-fields relating to *Function*, *Catalytic activity*, *Cofactor*, *Pathway*, *Developmental stage*, *Subcellular location* and *Disease*. Each block is headed by up to three IDs to clarify to which family the annotation belongs. However, as the aim is to provide super-family level annotation, relevant structure and similarity information is also gathered: e.g. as

GPCRs share a 7TM helical bundle architecture, it seems reasonable to use the structure of rhodopsin as a template for all super-family members. Information from the Subunit, Domain and Similarity sub-fields is therefore collated and assessed to determine whether it is common to all super-family members, and annotation is inherited only if it is shared by a majority. This information is provided after the family-specific annotation blocks, under the heading 'Superfamily and structural annotation', to indicate that it relates to all families.

For domains, PRECIS again takes the five most populous families (headed by up to three family members) and digests the same CC sub-fields. Now, however, common information from the *Subunit*, *Domain* and *Similarity* sub-fields in each family is included within the relevant annotation block, as this will be pertinent to the specific protein family rather than to the entire domain family: e.g. the structures of tyrosine protein kinase, phosphotransferase, growth factor receptor adaptor protein, etc., are different, despite including one or more SH2 domains—the structure of one of these proteins could therefore not serve as a template for all SH2-containing proteins. Ideally, we would indicate the structure of the SH2 domain itself, but there is often no simple way to ascertain this directly, because SWISS-PROT deals with sequence—rather than domain-family-specific information.

In summary, the final report offers structured common, shared or special information tailored to whether a family, super-family or domain is being processed. For domains and super-families, we confine the report to the five most populous families for pragmatic reasons—some domains and super-families can include thousands of sequences from dozens of families, and the reports would become too long were we to include representative information from all of them. However, all information excluded from the report is retained for further possible processing.

IMPLEMENTATION

PRECIS is written in Perl, which allows rapid parsing of semi-structured text-based data. The software can analyse hundreds of SWISS-PROT entries in minutes on a desktop PC.

APPLICATIONS

For protein family database curators, the manual burdens inherent in the process of annotation are large. Typically, this might involve: conducting a database search to identify matches to a query sequence; examining the matches individually to determine their biological significance; tracing relevant annotation (e.g. from SWISS-PROT, or from abstracts, full texts, etc.) and ultimately, deriving consensus annotation from scores of representatives from the matched set. Clearly, PRECIS can ease some of the more laborious of these annotation tasks. In its first application, we therefore used it to add annotation to an automatic supplement to PRINTS, termed prePRINTS. This allows automatically generated fingerprints to have at least some level of annotation associated with them so that users do not reach dead-ends when they find matches in prePRINTS (a problem with unannotated hidden Markov models and pre-profiles).

To make the program more generally useful to biologists who do not have access to our fingerprint software, in a second application, we tried to simplify the process of gathering sets of SWISS-PROT IDs. As BLAST has become a familiar search tool for biologists, we chose it as our starting point. For each match to a query sequence, BLAST output is characterized by database IDs/AC numbers and sequence descriptions, plus scores and *E*-values. Matches that cluster at the top of hitlists with low *E*-values are likely to be related. However, to discover more about the biology of the matched proteins, the individual database entries must be visited, read and digested one at a time. We therefore configured PRECIS to do this automatically. The system accepts a single sequence in FASTA format, BLAST is executed, and SWISS-PROT IDs found below a user-defined *E*-value cut-off are extracted. PRECIS then distils relevant information into a formatted report, as illustrated in Figure 4. The BLAST hitlist can subsequently be refined at will to include or preclude matches from the annotation process.

To make further information accessible to augment the reports, hyperlinks are provided to SWISS-PROT, to family and structure databases, and to the online abstracts of the literature references included in the final report. It is thus possible to take a single sequence and amass a wealth of data, including a formatted family report, the complete data-set from which the report was derived, and relevant abstracts and/or reprints from online literature sources.

RESULTS

Example output from PRECIS is shown in Figure 4, which illustrates the report generated for the rhodopsin family. BLASTing SWISS-PROT with TrEMBL entry Q9PUZ5, a triggerfish rhodopsin, returned 107 significant matches (*E*-values $<$ 1e–60). For each ID, the full entry was retrieved from SWISS-PROT. The complete data-set, occupying 204 pages of text, was analysed and processed, and the results distilled into a 1.5-page report.

The title of the report is 'Rhodopsin', which was the most frequent description (DE) occurring in 84 out of 107 SWISS-PROT entries (cf. Fig. 1). Database cross-references are provided for PRINTS, PROSITE, Pfam, InterPro, PDB, SCOP, CATH and MIM (nb. because a reference to PDB was found, links to SCOP and CATH have been synthesized). Seven literature references are given, the last three relating to structure determinations.

In the body of the annotation are descriptions of the protein function and its associated diseases; each disease paragraph is assigned its relevant ID so that the information can be traced

```
Rhodopsin

PRINTS; PR00237 GPCRRHODOPSN; PR00238 OPSIN; PR00579 RHODOPSIN
PROSITE; PS00237 G_PROTEIN_RECEP_F1_1; PS00238 OPSIN; PS50262 G_PROTEIN_RECEP_F1_2
PFAM; PF00001 7tm_1
INTERPRO; IPR000276; IPR001760
PDB; 1BOJ; 1BOK; 1EDS; 1EDV; 1EDW; 1EDX; 1F88; 1FDF
SCOP; 1BOJ; 1BOK; 1EDS; 1EDV; 1EDW; 1EDX; 1F88; 1FDF
CATH; 1BOJ; 1BOK; 1EDS; 1EDV; 1EDW; 1EDX; 1F88; 1FDF
MIM; 163500; 180380; 268000

1. FYHRQUIST, N., DONNER, K., HARGRAVE, P.A., MCDOWELL, J.H., POPP, M.P. AND
SMITH, W.C.
Rhodopsins from three frog and toad species: sequences and functional comparisons.
EXP.EYE RES. 66 295-305 (1998).

2. ARCHER, S.N. AND HIRANO, J.
Opsin sequences of the rod visual pigments in two species of Poeciliid fish.
J.FISH BIOL. 51 215-219 (1997).

3. HUNT, D.M., FITZGIBBON, J., SLOBODYANYUK, S.J., BOWMAKER, J.K. AND DULAI,
K.S.
Molecular evolution of the cottoid fish endemic to Lake Baikal deduced from nuclear DNA evidence.
MOL.PHYLOGENET.EVOL. 8 415-422 (1997).

4. GALE, J.M., TOBEY, R.A., D'ANNA, A.
Localization and DNA sequence of a replication origin in the rhodopsin gene locus of Chinese
hamster cells.
J.MOL.BIOL. 224 343-358 (1992).

5. PALCZEWSKI, K., KUMASAKA, T., HORI, T., BEHNKE, C.A., MOTOSHIMA, H., FOX,
B.A., LE TRONG, I., TELLER, D.C., OKADA, T., STENKAMP, R.E., YAMAMOTO, M. AND
MIYANO, M.
Crystal structure of rhodopsin: a G protein-coupled receptor.
SCIENCE 289 739-745 (2000).

6. YEAGLE, P.L., ALDERFER, J.L. AND ALBERT, A.D.
Structure of the third cytoplasmic loop of bovine rhodopsin.
BIOCHEMISTRY 34 14621-14625 (1995).

7. YEAGLE, P.L., SALLOUM, A., CHOPRA, A., BHAWSAR, N., ALI, L., KUZMANOVSKI,
G., ALDERFER, J.L. AND ALBERT, A.D.
Structures of the intradiskal loops and amino terminus of the G-protein receptor, rhodopsin.
J.PEPT.RES. 55 455-465 (2000).

Function:
Visual pigments are the light-absorbing molecules that mediate vision. They consist of an
apoprotein, opsin, covalently linked to cis-retinal.

Additional Info:
Integral membrane protein.

Disease:
Defects in rho are one of the causes of autosomal dominant retinitis pigmentosa (adrp). Patients
typically have night vision blindness and loss of midperipheral visual field; as their condition
progresses, they lose their far peripheral visual field and eventually central vision as well.
(OPSD_HUMAN).

Defects in rho are one of the causes of autosomal recessive retinitis pigmentosa (arrp).
(OPSD_HUMAN).

Defects in rho are also one of the causes of congenital stationary night blindness (csnb4).
(OPSD_HUMAN).

Family and structural information:
The structure has been determined, e.g., "Crystal structure of rhodopsin: a G
protein-coupled receptor" [5]; "Structure of the third cytoplasmic loop of
bovine rhodopsin" [6]; "Structures of the intradiskal loops and amino terminus
of the G-protein receptor, rhodopsin" [7].

Belongs to family 1 of g-protein coupled receptors. Opsin subfamily.

Keywords: Photoreceptor; Retinal protein; Transmembrane; Glycoprotein; Vision;
Phosphorylation; Lipoprotein; Palmitate; G-protein coupled receptor; Acetylation; 3D-structure;
Retinitis pigmentosa; Disease mutation.
```

Fig. 4. Example PRECIS output for the rhodopsin family.

and placed in the correct species context. Note that although the SWISS-PROT entry in Figure 1 contains a CC *Developmental Stage* sub-field, this does not appear in the report as it does not occur in a majority of IDs (2/107) and is thus likely to be sequence-rather than family-specific. Finally, PRECIS reports that the structure of the protein has been determined and provides pointers to the appropriate literature references. The report then indicates that the sequences belong to the opsin subfamily of family 1 of GPCRs, and provides a set of keywords.

DISCUSSION

There have been several endeavours to address the problem of annotating sequence data automatically, but this is a non-trivial computational task and few user-friendly tools have emerged. To support researchers in the process of annotating newly determined sequences, we therefore developed PRECIS, which digests information in related SWISS-PROT entries. The decision not to use the on-line literature was a pragmatic one, based on several important observations: (i) MEDLINE abstracts are not structured, making information retrieval difficult; (ii) typically, abstracts are too short to provide informative, generic information on protein families and (iii) access to full texts on the Internet is still fairly limited. However, although not yet used by the system, PRECIS currently collects all MEDLINE abstracts, which we plan to use later to augment the core annotation extracted from SWISS-PROT.

Using SWISS-PROT allows us to exploit both the in-built structure of its entries and the richness of information already incorporated by teams of annotators. But this approach is clearly limited by the quality and extent of annotation available. If there is little or no existing annotation, PRECIS will at best provide some literature references, database cross-references and keywords; and, if there are consistent errors in SWISS-PROT, PRECIS will inherit them. However, literature and database cross-references are significantly more useful to annotators than mere lists of keywords, making the retrieval of further information relatively straightforward. The approach has the further advantage that random errors will be filtered out, as the program uses the weight of evidence gathered from multiple sequence entries to draw its conclusions.

To render PRECIS useful to the wider biological community, we implemented an online system capable of exploiting BLAST to generate a structured protein family report from a single sequence. The system requires only the initial query sequence and an *E*-value cut-off, above which sequences are not permitted into the annotation-culling process.

The reports are English-like, in the sense that they largely re-use existing human annotation, but consequently exhibit the rather clipped, note-like style typical of SWISS-PROT. Although informative, the result is inevitably not the same as would be produced by an annotator working from scratch with a diverse range of information sources. Nevertheless, PRECIS is a step forward in the development of automatic annotation tools. It reaches beyond the tool developed to annotate TrEMBL automatically (Bairoch and Apweiler, 2000)—annotation here refers to database cross-references and similarity assignments (based on sequence matches in resources such as PROSITE, PRINTS, etc.), or predicted features of the sequence, such as subcellular location, transmembrane domains, and so on (Moeller *et al*., 1999). It also makes a significant advance over rather simplistic tools that merely generate keyword lists. For the database curator, such tools are useful for characterizing unknown sequences, but do little to help with the more onerous task of writing useful annotation.

FUTURE DIRECTIONS

One future development will be to incorporate a formally structured meta-data layer into the system. PRECIS was developed as a proof-of-concept, to see if informative reports could be distilled from sets of SWISS-PROT entries. However, to be maximally useful, the tool should be able to generate information in a form that is both human readable and machine parseable.

We are also re-visiting the literature, exploring text-mining and natural language processing methods to extract further useful information to assist the annotation-gathering process, and to facilitate the addition of more extensive annotation during the migration of entries in prePRINTS to PRINTS. These techniques should also help to address issues of editorial/quality control. Although we strive to produce non-redundant reports, PRECIS sometimes generates duplications. Where descriptions retrieved from CC sub-fields are identical, duplicates can be easily removed. More difficult to tackle are cases where descriptions show small but potentially important variations: e.g. “Primary transducing effect is inhibition of adenylate cyclase” and “Primary transducing effect is adenylate cyclase inhibition”, although identical in biological meaning, have syntactical differences that are difficult to resolve automatically. In the meantime, it is safer to include at least some level of redundancy, which a human curator can remove later.

CONCLUSIONS

Tools to assist and to automate the process of annotation are sorely needed. To address this need, we have developed PRECIS, which generates protein reports from related SWISS-PROT entries. For ease of use, we have made the program available online and coupled it to BLAST, necessitating only a single sequence as input. The software, although dependent on SWISS-PROT, represents a useful first step towards producing (a) a fully automatic annotation tool and (b) a decision-support framework that a human annotator can use to gather more detailed information. The tool

has considerable potential to assist curators of protein family databases, where it would have the dual advantage of reducing current manual burdens, and of creating information in a format that is consistent, computer readable and readily updateable.

ACKNOWLEDGEMENTS

We are grateful to the BBSRC for support (AM and JR). TKA is a Royal Society University Research Fellow.

REFERENCES

Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**, 403–410.

Andrade,M.A. and Valencia,A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.

Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Croning,M.D.R., Durbin,R. *et al.* (2001) InterPro—An integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.

Attwood,T.K., Blythe,M., Flower,D.R., Gaulton,A., Mabey,J.E., Maudling,N., McGregor,L., Mitchell,A., Moulton,G., Paine,K. and Scordis,P. (2002) PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res.*, **30**, 239–241.

Bairoch,A. (2000) The ENZYME data bank in 2000. *Nucleic Acids Res.*, **28**, 304–305.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J, Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.

Frishman,D. and Mewes,H.-W. (1997) PEDANTic genome analysis. *Trends Genet.*, **13**, 415–416.

Gaasterland,T. and Sensen,C.W. (1996) Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie*, **78**, 302–310.

Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.

Holm,L. and Sander,C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, **27**, 244–247.

Lo Conte,L., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.

Moeller,S., Leser,U., Fleischmann,W. and Apweiler,R. (1999) EDITtoTrEMBL: a distributed approach to high-quality automated protein sequence annotation. *Bioinformatics*, **15**, 219–227.

Pearl,F.M.G., Martin,N., Bray,J.E., Buchan,D.W.A., Harrison,A.P., Lee,D., Reeves,G.R., Shepherd,A.J., Sillitoe,I., Todd,A.E. *et al.* (2001) A rapid classification protocol for the CATH Domain Database to support structural genomics. *Nucleic Acids Res.*, **29**, 223–227.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA.*, **85**, 2444–2448.

Scharf,M., Schneider,R., Casari,G., Bork,P., Valencia,A., Ouzounis,C. and Sander,C. (1994) GeneQuiz: a workbench for sequence analysis. In R. Altman, D. Brutlag, P. Karp, R. Lathrop and D. Searls (eds), *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 348–353.

Selley,J.N., Swift,J. and Attwood,T.K. (2001) EASY—an Expert Analysis SYstem for interpreting database search outputs. *Bioinformatics*, **17**, 105–106.

Wise,M.J. (2000) Protein annotator's assistant. *Trends Biochem. Sci.*, **25**, 252–253.